LA-UR-16-24592
Approved for public release; distribution is unlimited.

Title: Alternatives to accuracy and bias metrics based on percentage errors for radiation belt modeling applications

Author(s): Morley, Steven Karl

Intended for: Report

Issued: 2016-07-01 (rev.1)

Alternatives to accuracy and bias metrics based on percentage errors for radiation belt modeling applications

S. K. Morley[1]

[1] *Los Alamos National Laboratory, New Mexico, USA*

28th June 2016

Abstract

This report reviews existing literature describing forecast accuracy metrics, concentrating on those based on relative errors and percentage errors. We then review how the most common of these metrics, the mean absolute percentage error (MAPE), has been applied in recent radiation belt modeling literature. Finally, we describe metrics based on the ratios of predicted to observed values (the accuracy ratio) that address the drawbacks inherent in using MAPE. Specifically we define and recommend the median log accuracy ratio as a measure of bias and the median symmetric accuracy as a measure of accuracy.

1 Introduction

The utility, or value, of any forecast model is determined by how well the forecast predicts the quantities being modeled. There exists, however, a wide range of metrics to assess forecast quality and a similarly wide range of views on just what a "good" forecast is. One key measure of the quality of a forecast is in how much it deviates from the observation and that is what will be discussed in this report. We begin by briefly introducing some quantitative attributes of forecast performance, followed by definitions of metrics to quantitatively describe these attributes in the case of continuous predictands.

Although a forecast is strictly a prediction of events that have not yet occurred, this report treats simulation results as a forecast, regardless of the time interval.

1.1 Forecast performance and metrics

Wilks [2006, (section 7.1.3)] gives a "partial list of scalar aspects, or attributes, of forecast quality" with six descriptors of forecast performance. Though all of these are undoubtedly important in characterizing the quality of a forecast, this report will address only the first two in any detail: accuracy, and bias.

Scalar accuracy measures describe the average error between a given forecast and the corresponding observations. Various metrics can be used for this (e.g., mean squared error) and a selection will be described in the next section [see, e.g., *Walther and Moore*, 2005; *Hyndman and Koehler*, 2006; *Wilks*, 2006]. Our discussion begins with the cornerstone of most metrics of accuracy and bias: the forecast error, ε

$$\varepsilon = y - x \tag{1}$$

where x denotes the observation and y denotes the predicted value. To illustrate the concept we predict the magnitude of Earth's magnetic field at two locations to be [48,57] nT, but the observed values are [50,61] nT. The forecast error is then [−2, −4] nT.

When we have multiple pairs of forecast and observation it is helpful to aggregate these errors and present summary statistics. The bias, or systematic bias, describes the difference between the average forecast and the average observation. A standard measure of forecast bias is the mean error (ME), defined as:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon) \tag{2}$$

where n is the number of observation-forecast pairs, and the subscript i denotes the i^{th} element of the series. In our previous example the forecast values were consistent underestimates of the observed value. Forecasts that, on average, over- or under-estimate the observed value display bias. Calculating the ME for the example above we have $((-2)+(-4))/2 = -3$ nT. A negative number indicates a systematic under-prediction, whereas a positive bias would indicate a systematic over-prediction.

For data that have different scales, scale-independent accuracy measures are often recommended. Although the variability in electron fluxes at a given location and energy can be large, scale-dependent measures would still be appropriate. However, there can be several orders of magnitude difference between electron fluxes at L$\simeq$ 4 and geosynchronous orbit, with each location displaying different levels of variability. Thus comparing scale-dependent accuracy measures can be problematic. Similarly, the measurements across a single orbit of a satellite in a highly-elliptical orbit cover regions that could be argued to be of different scale and dynamics.

One approach to giving more equal weight to errors across several orders of magnitude is to use metrics that are based on relative errors (including percentage errors) or are otherwised scaled to normalize the errors. Alternatively, the data themselves can be transformed through the application of a power function, such as taking logarithms or applying a Box-Cox transform [*Wilks*, 2006]. By transforming the data this way, the use of scale-dependent accuracy measures may be better justified, as well as application of methods that assume homoscedasticity [*Sheskin*, 2007]. We note, however, that transforming the data alters the scale and may invalidate the assumptions behind other analyses. We will first introduce scale-dependent metrics, followed by two classes of scale-independent metrics. We will then focus more closely on the mean absolute percentage error and its use in recent literature, before proposing a new accuracy measure based on relative errors.

1.2 Metrics based on scale-dependent errors

In some works *scale-dependent* errors are referred to as absolute errors [e.g. *Walther and Moore*, 2005], but we avoid that terminology to avoid confusion between an absolute measure (in the sense that it has not been scaled) and the absolute value of a measure (i.e., $|x|$).

Like the bias, accuracy measures typically begin with the forecast errors, ε_i, but then transform the data so that the direction of difference is removed. This is typically done by either squaring the forecast error or taking the absolute value (modulus) of the forecast error. The mean squared error (MSE) takes the former approach.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\varepsilon^2 \tag{3}$$

It can be seen that the mean squared error is analogous to the variance, and penalizes large errors more heavily than small errors. When fitting a regression model, use of ordinary least squares (OLS)

minimizes this metric (though *forecast error* and *fitting error* have opposite sign conventions). Other error measures can be used for robust regression, but typically require iterative, numerical minimization strategies [*Wilks*, 2006].

Squaring the errors leads to the units and scale being different from the forecast quantity, which makes the MSE difficult to interpret. Transforming MSE back to the original scale by taking the square root then gives the root mean squared error (RMSE):

$$RMSE = \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon^2 \right)^{\frac{1}{2}} \tag{4}$$

As we are concerned with estimating the accuracy of a forecast, which will likely not be derived from an OLS regression model, the decision of which error metric should be used depends on the relative cost of different errors. Two pertinent questions here are:

1. If the error doubles, is this twice as bad? Or is it more than twice as bad?
2. Is an overestimate worse than an underestimate of the same magnitude?

The two questions can be equivalently phrased as "What is the form of the cost function (e.g., linear or quadratic)?", and "Is the cost function symmetric?" If we wish to reduce the penalty on large errors – i.e., use a linear loss function, rather than a quadratic loss function – we can use the mean absolute error (MAE). This is defined as

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\varepsilon| \tag{5}$$

This metric is more resistant to outliers, as it uses $|\varepsilon|$ rather than ε^2. It may, therefore, be more appropriate in cases where the errors are not normally distributed or where large forecast errors are not required to be weighted more heavily.

Both the MSE and MAE estimate the *location* (central tendency) of the error distribution using the mean. As the mean is not a robust measure, we can improve the robustness of our accuracy metric by using a common robust measure of location: the median. Replacing the mean function in equation 5 with the median function (M) gives us the median absolute error (MdAE).

$$MdAE = M\left(|\varepsilon|\right) \tag{6}$$

A good summary of unscaled measures of accuracy and bias can be found in *Walther and Moore* [2005]. We note here that unscaled metrics imply that deviations of the same magnitude have equal importance at different magnitudes of the base quantity. For example, an error of $\varepsilon = 100$ is penalized equally at $x = 10^3$ and $x = 10^6$.

1.3 Metrics based on relative and percentage errors

When measuring the accuracy of a prediction the magnitude of relative error (MRE) is often used, it is defined as the absolute value of the ratio of the error to the actual observed value. When multiplied by 100 this gives the absolute percentage error (APE). This measure is generally only used when the quantity of interest is strictly positive, and we shall make this assumption throughout.

We begin by defining the relative error, η:

$$\eta = \frac{y - x}{x} = \frac{\varepsilon}{x} \tag{7}$$

Following the discussion given in section 1.2 we then remove the direction of difference by taking $|\eta|$, the absolute relative error. As noted previously, to assess the accuracy of multiple predictions a way of aggregating is required. If we adopt the arithmetic mean then we shall be using the mean absolute percentage error (MAPE). In some disciplines this is known as the "mean magnitude of relative error" (MMRE) [*Kitchenham et al.*, 2001]. Defining relative error with equation 7, the magnitude of relative error is therefore:

$$MRE = \left| \frac{y - x}{x} \right| = |\eta| \tag{8}$$

and the mean absolute percentage error is then:

$$MAPE = 100 \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - x_i}{x_i} \right| \tag{9}$$

$$= 100 \frac{1}{n} \sum_{i=1}^{n} |\eta_i| \tag{10}$$

We note here that relative and percentage error metrics imply that deviations of the same order have equal importance at different magnitudes of the base quantity. For example, an error of $\varepsilon = 100$ where $x = 10^3$ has an equal penalty to an error $\varepsilon = 1$ where $x = 10$ – both give a relative error of 0.1, and thus a percentage error of 10%.

MAPE is used across many different fields of research, from population research [e.g. *Swanson et al.*, 2000] to business forecasting [e.g. *Kohzadi et al.*, 1996] and atmospheric science [e.g. *Grillakis et al.*, 2013; *Zheng and Rosenfeld*, 2015]. MAPE has also been used in validation of radiation belt models [*Kim et al.*, 2012; *Tu et al.*, 2013; *Li et al.*, 2014], and these are discussed further in section 2. However, MAPE is not without problems that may be pertinent for radiation belt forecasts. The following problems have been noted by various authors:

1. MAPE becomes undefined when the true value is zero. [*Hyndman and Koehler*, 2006]
2. MAPE is asymmetric with respect to over- and under-forecasting. [*Makridakis*, 1993; *Hyndman and Koehler*, 2006; *Tofallis*, 2015]
3. APE is constrained to be positive, so its distribution is generally positively skewed. [*Swanson et al.*, 2000; *Hyndman and Koehler*, 2006]
4. MAPE is not resistant to outliers [*Swanson et al.*, 2000; *Tofallis*, 2015].

We note that MAPE is not an appropriate metric where the quantity being modeled can be zero. Indeed, *Tofallis* [2015] note that APE "is generally only used when the quantity of interest is strictly positive". We also note that unless the data used are ratio-level data [*Sheskin*, 2007], the APE has limited meaning [*Hyndman and Anathasopoulos*, 2014]. For example, radiation belt fluxes are constrained to be positive and the units of flux have a true zero (which, practically speaking, is unlikely to be encountered), therefore APE can be used for radiation belt flux predictions.

To elaborate on the second point, a prediction of 1000 where the observed value is 500 gives a different magnitude of error (100%) than a prediction of 500 where the observed value is 1000 (50%). Under-prediction is therefore less heavily penalized that over-prediction, even if the magnitude of the error is the same.

Given that APE have a lower bound of zero but have no upper bound, they are likely to be skewed positive. Take a case where the forecast errors are distributed approximately normally, and

are symmetric about the true value. The distribution of APE is now highly skewed; by using the arithmetic mean, which is a poor measure of central tendency in skewed distributions, MAPE is prone to overstating the error.

Finally, MAPE is easily affected by outliers as the mean has a breakdown point of zero. Given a set of predictions with APE of [5,3,10,2,5,120] %, MAPE takes the value 24.16%; reducing the error on the final prediction from 120% to 30% reduces the MAPE to 9.17%. Therefore any large errors due to bad data, late (or early) prediction of a large change, etc. will all be heavily penalized by taking the arithmetic mean.

1.4 Metrics based on scaled errors

Hyndman and Koehler [2006] proposed an alternative to percentage errors, called *scaled errors*, and constructed accuracy metrics based on the scaled error. Their original definition of the scaled error, q, is:

$$q_i = \frac{x_i - y_i}{\frac{1}{n-1}\sum_{i=2}^{n} |x_i - x_{i-1}|} \tag{11}$$

where the error is scaled based on the error expected from a persistence forecast. The benefit of this scaling is that a scaled error is below 1 if the forecast outperforms the average error from the persistence forecast. This method was developed for time series data, but is not appropriate where the observing location changes with time (such as flux data from a satellite in a highly elliptical orbit). In this case, the appropriate forecast to scale by might be "orbital persistence"; assuming the orbital characteristics change slowly then on successive orbits the satellite will return to approximately the same location and the value at that location is then taken instead of x_{i-1}. A further modification of the scaled error was given by *Hyndman and Anathasopoulos* [2014], where the scaling is by comparison with the error compared to the mean forecast. In the case of a satellite covering a wide range of locations the mean forecast is also inappropriate. However, the error relative to a climatological mean for the current location of the satellite provides a scaling that is meaningful.

$$q_i = \frac{x_i - y_i}{\frac{1}{n-1}\sum_{i=2}^{n} |x_i - c_i|} \tag{12}$$

where c_i is the climatological prediction for the location of measurement x_i. The mean absolute scaled error (MASE) is then given by

$$MASE = \frac{1}{n}\sum_{i=1}^{n} |q_i| \tag{13}$$

It can be seen that this is essentially the mean absolute error of the model, normalized by the mean absolute error of a benchmark (here a climatological model). When computing MASE for a number of models, we note that if a climatological model is used then a MASE of exactly 1 will result. The scaling is also the same for all models being compared as the data and the climatological values will not change, only the forecast values change with each different model. Though this metric makes comparison of models intuitive, this metric suffers from the actual values being difficult to interpret as a magnitude of error; the ratio of mean absolute error to mean absolute error of a benchmark gives little direct information about the size of the errors in the specific model. For this reason, MAPE remains very popular and we will discuss its use in radiation belt studies before proposing an alternative that addresses some of its drawbacks.

2 MAPE in magnetospheric modeling

As noted previously, MAPE has been used directly in assessing the accuracy of radiation belt models. This application provided the motivation for this report, and hence we examine how MAPE has been used in the literature, but we also briefly consider some related prediction studies.

The formulation given by *Kim et al.* [2012] [see also *Tu et al.*, 2013; *Li et al.*, 2014] uses the logarithm of the predicted (and observed) quantity in equation 9, although these authors predict flux (the subsequent works cited use phase space density), not their logs. The MAPE presented therefore cannot be interpreted intuitively as the percentage reported is not in the quantity being modeled, does not have physical units, nor is it trivial to understand this as a relative error. For example, given a constant measured flux of $10^5\,\mathrm{cm^{-2}s^{-1}sr^{-1}keV^{-1}}$ and a constant prediction of $1.7\times10^5\,\mathrm{cm^{-2}s^{-1}sr^{-1}keV^{-1}}$ we can see that the prediction is consistently a factor of 1.7 (70%) from the measurement. If we calculate the MAPE on the untransformed data we get:

$$MAPE = 100\left|\frac{1.7\times10^5-1\times10^5}{1\times10^5}\right| \quad (14)$$

$$= 70\% \quad (15)$$

which gives the intuitive result of a 70% error. When we log-transform the data and repeat this process we find:

$$MAPE = 100\left|\frac{\log(1.7\times10^5)-\log(1\times10^5)}{\log(1\times10^5)}\right| \quad (16)$$

$$= 100\left|\frac{\log(1.7\times10^5/1\times10^5)}{\log(1\times10^5)}\right| \quad (17)$$

$$\simeq 4.6\% \quad (18)$$

Note that by taking logarithms we now normalize the log of the flux ratio by the log of the measured flux, thereby breaking the interpretation of MAPE with respect to the predicted quantity. This quantity now varies with the magnitude of the observation. If our prediction-observation pair is (1.7×10^5, 10^5) or (1.7×10^2, 10^2) then calculating MAPE on untransformed data gives 70% in both cases. Calculating the MAPE using the log of these data gives 4.6% and 11.5%. Put another way, by log-transforming the data we no longer strictly have ratio-level data, MAPE no longer has an intuitive meaning, and MAPE is no longer an appropriate metric.

To predict the effective dose of galactic cosmic radiation received on trans-polar aviation routes, *Hwang et al.* [2015] developed a model that forecasts the heliocentric potential (HCP) from a lagged time-series of monthly sunspot number. The HCP is a required input for the Federal Aviation Administration's CARI-6M software for dose estimation. The modeled HCP presented by *Hwang et al.* [2015] shows less variability than the observed HCP, with a tendency for the low values to be slightly overpredicted and the high values to be significantly underpredicted. Since MAPE more heavily penalizes the overprediction, it is possible that the accuracy reported by *Hwang et al.* [2015] overstates the true accuracy of the model.

Zhelavskaya et al. [2016] have developed a neural network to predict the frequency of the upper-hybrid resonance to derive electron number densities in the inner magnetosphere, using Van Allen Probes electric field data. These authors used MAPE to assess the accuracy of their predictions, both in predicted frequency and predicted number density. We note that the electron number density, like radiation belt electron flux, is constrained to be positive and has a physically meaningful zero. Further, the electron number density can vary by orders of magnitude over a single orbit as well as at a fixed location due to dynamical processes.

Hwang et al. [2015] and *Zhelavskaya et al.* [2016] calculated MAPE directly, without first transforming the data, and their reported percentage errors are therefore directly interpretable, though still suffer from the drawbacks described in section 1.3.

3 Robust, symmetric measures of bias and accuracy

We begin by defining the accuracy ratio, Q, as y/x; that is, the ratio of the predicted value to the observed value [1]. The name "accuracy ratio" was coined by *Tofallis* [2015], who note that Q is the complement of the relative error ($Q = 1 - \eta$) and so will have the same distribution as the relative error, but shifted by one unit. *Tofallis* [2015] also showed that Q is a superior accuracy measure to MAPE for heteroscedastic data (as is often the case with space physics data, such as radiation belt electron fluxes [e.g. *Reeves et al.*, 2011; *Morley et al.*, 2016]).

We note that previous work on energetic electron data, in both fluxes and phase space densities, have used ratios of the observed to predicted values. *Chen et al.* [2007] defined the "PSD matching ratio", R, [see also *Yu et al.*, 2014] as the ratio of phase space densities, where the denominator is always the smaller of the two values.

$$R = \frac{f_{large}}{f_{small}} \tag{19}$$

Morley et al. [2016] used the accuracy ratio to compare electron fluxes computed from the Global Positioning System constellation with "gold standard" measurements from the Van Allen Probes mission. When presenting summaries of these data as boxplots, *Morley et al.* [2016] showed $\log_{10}(Q)$ "so that the ratios are symmetric both above and below 1." Taking the logarithm ensures that a factor of 3 difference between x and y is the same magnitude of error, regardless of the direction of error. However, even though log-transforming the data will tend to symmetrize positively skewed distribution, the actual distributions of $\log(Q)$ may not be symmetric. For this reason, *Morley et al.* [2016] used the median of $\log(Q)$ (MdLQ) as a measure of central tendency

$$MdLQ = M\left(\log(Q_i)\right) \tag{20}$$

this quantity also represents a robust measure of bias, though it suffers from a lack of intuitive interpretability.

3.1 Accuracy

We here propose a measure of accuracy that use logarithms of the accuracy ratio, thereby mitigating many of the problems inherent in using MAPE, but that maintain the interpretability of MAPE. Specifically, we follow the lead of *Tofallis* [2015] and *Morley et al.* [2016] in using $\log(Q)$, but modify our accuracy metric such that it is interpretable as a percentage error.

We begin by taking the absolute values of $\log(Q)$, then exponentiating to return it to the original units and scale. This transformation ensures that the metric is symmetric in the sense that switching the values of the predicted and observed value give the same error (unlike MAPE).

$$R = \exp(|\log(Q)|) \tag{21}$$

This can also be seen to be the "matching ratio" of *Chen et al.* [2007].

[1] *Kitchenham et al.* [2001] present an interesting discussion of the relationship of Q to the mean absolute error and recommend this as a measure of prediction accuracy

The symmetric accuracy, aggregated over a set of forecast-observation pairs using the median and transformed to a percentage scale, is then given as:

$$\zeta = 100 \left(\exp \left(M \left(\left| \log(Q_i) \right| \right) \right) - 1 \right) \tag{22}$$

where M denotes the median function. The subtraction of 1 changes the lower bound from unity to zero, allowing the interpretation as a (symmetric) fractional error, and multiplying this by 100 yields an equivalent percentage error. This metric, ζ, is therefore named the median symmetric accuracy. If we now return to our example from section 2 we can see that for two prediction-observation pairs, $(1.7 \times 10^5,\ 10^5)$ and $(1.7 \times 10^2,\ 10^2)$, ζ is 70% in both cases; this is the same as the correct application of MAPE. This result is also symmetric with respect to the reversal of the predictions and observations, in contrast with MAPE.

3.2 Bias

As we take the absolute values of $\log(Q)$ we lose information about systematic bias in the prediction. Therefore, by removing the modulus (and the transformation to a percentage scale) we recover a measure of bias: the median accuracy ratio, β_M

$$\begin{aligned} \beta_M &= \exp(M \left(\log(Q_i) \right) & (23) \\ &= M \left(Q_i \right) & (24) \end{aligned}$$

As the median is a rank order statistic it is invariant with respect to the log-transform and the exponentiation, thus we can simplify this to equation 24 [cf. *Morley et al.*, 2016]. If we instead consider a different location function, this simplification may not hold. If the distribution of $\log(Q)$ is reasonably symmetric then equation 24 will approximate the arithmetic mean of $\log(Q)$. If we substitute the median function with the mean function (μ), we recover the geometric mean of Q. We can therefore define a related metric, the geometric mean accuracy ratio, β_μ:

$$\begin{aligned} \beta_\mu &= \exp \left(\mu \left(\log(Q_i) \right) \right) & (25) \\ &= \left(\prod_{i=1}^{n} Q_i \right)^{\frac{1}{n}} & (26) \end{aligned}$$

Either of these measures will give values smaller than 1 for a systematic underprediction, and values greater than 1 for a systematic overprediction. Although β_μ has a clearer mathematical origin, we prefer β_M as it is more resistant to outliers and is more intuitive, due to its use of the median. The physical meaning of the accuracy ratio is also clear, making the median accuracy ratio an easily interpretable quantity. It should be noted that β_M is not symmetric about 1. If symmetry about 1 is required then the final exponentiation should not be taken, leaving us with the median log accuracy ratio, as plotted by *Morley et al.* [2016] and given in equation 20. This final metric has the benefit that underprediction will give a negative value of $M(\log(Q))$ and over-prediction will give a positive value; an unbiased forecast will yield $M(\log(Q)) = 0$. This symmetry about zero then mirrors the more common measure of bias, the mean error. For this reason, we recommend $M(\log(Q))$ as a measure of bias. The choice of base will determine the level of interpretability for any given data set.

4 Summary

We have introduced a number of commonly-used forecast metrics, including the mean absolute percentage error (MAPE). A literature review has revealed a number of known problems with the

use of MAPE as a measure of forecast accuracy, with some applications in space physics leading to MAPE losing its most desirable quality: interpretability.

To address the drawbacks associated with MAPE while still preserving the interpretability we have expanded on the work of *Tofallis* [2015] and *Morley et al.* [2016] to develop an accuracy measure based on the logarithm of the accuracy ratio. This measure can be interpreted as a percentage error, but does not penalize over-and under-prediction differently. We call this metric the "median symmetric accuracy", ζ, which is defined as

$$\zeta = 100\left(\exp\left(M\left(\left|\log\left(\frac{y_i}{x_i}\right)\right|\right)\right) - 1\right)$$

To indicate bias in a symmetric manner we recommend the median log accuracy ratio (MdLQ)

$$MdLQ = M\left(\log\left(\frac{y_i}{x_i}\right)\right)$$

as used by *Morley et al.* [2016], as it can be interpreted similarly to the mean error where negative values indicate a systematic under-prediction and positive values indicate a systematic overprediction. By using base 10 logarithms an order of magnitude difference is given by MdLQ=1, and a factor of 2 difference is given by MdLQ$\simeq$0.3.

References

Chen, Y., R. H. W. Friedel, G. D. Reeves, T. E. Cayton, and R. Christensen (2007), Multisatellite determination of the relativistic electron phase space density at geosynchronous orbit: An integrated investigation during geomagnetic storm times. *Journal of Geophysical Research: Space Physics*, *112*(A11), A11214, doi:10.1029/2007JA012314.

Grillakis, M. G., A. G. Koutroulis, and I. K. Tsanis (2013), Multisegment statistical bias correction of daily GCM precipitation output. *Journal of Geophysical Research: Atmospheres*, *118*(8), 3150–3162, doi:10.1002/jgrd.50323.

Hwang, J., K. C. Kim, K. Dokgo, E. Choi, and H. P. Kim (2015), Heliocentric potential (HCP) prediction model for nowscast of aviation radiation dose. *J. Astron. Space Sci.*, *22*(1), 39–44, doi:10.5140/JASS.2015.32.1.39.

Hyndman, R. J. and G. Anathasopoulos (2014), *Forecasting: Principles and Practice*. otexts.com, ISBN 978-0-9875071-0-5.

Hyndman, R. J. and A. B. Koehler (2006), Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679 – 688, doi:10.1016/j.ijforecast.2006.03.001.

Kim, K.-C., Y. Shprits, D. Subbotin, and B. Ni (2012), Relativistic radiation belt electron responses to GEM magnetic storms: Comparison of CRRES observations with 3-D VERB simulations. *Journal of Geophysical Research: Space Physics*, *117*(A8), A08221, doi:10.1029/2011JA017460.

Kitchenham, B. A., L. M. Pickard, S. G. MacDonell, and M. J. Shepperd (2001), What accuracy statistics really measure [software estimation]. *IEE Proceedings - Software*, *148*(3), 81–85, doi:10.1049/ip-sen:20010506.

Kohzadi, N., M. S. Boyd, B. Kermanshahi, and I. Kaastra (1996), A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing*, *10*(2), 169 – 181, doi:http://dx.doi.org/10.1016/0925-2312(95)00020-8, financial Applications, Part I.

Li, Z., M. Hudson, and Y. Chen (2014), Radial diffusion comparing a THEMIS statistical model with geosynchronous measurements as input. *Journal of Geophysical Research: Space Physics*, *119*(3), 1863–1873, doi:10.1002/2013JA019320.

Makridakis, S. (1993), Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, *9*(4), 527 – 529, doi:http://dx.doi.org/10.1016/0169-2070(93)90079-3.

Morley, S. K., J. P. Sullivan, M. G. Henderson, J. B. Blake, and D. N. Baker (2016), The Global Positioning System constellation as a space weather monitor: Comparison of electron measurements with Van Allen Probes data. *Space Weather*, *14*(2), 76–92, doi:10.1002/2015SW001339, 2015SW001339.

Reeves, G. D., S. K. Morley, R. H. W. Friedel, et al. (2011), On the relationship between relativistic electron flux and solar wind velocity: Paulikas and Blake revisited. *Journal of Geophysical Research: Space Physics*, *116*(A2), A02213, doi:10.1029/2010JA015735.

Sheskin, D. J. (2007), *Handbook of Parametric and Nonparametric Statistical Procedures, Fourth Edition*. Chapman and Hall/CRC.

Swanson, D. A., J. Tayman, and C. F. Barr (2000), A note on the measurement of accuracy for subnational demographic estimates. *Demography*, *37*(2), 193–201.

Tofallis, C. (2015), A better measure of relative prediction accuracy. *J. Oper. Res. Soc.*, *66*(8), 1352–1362.

Tu, W., G. S. Cunningham, Y. Chen, M. G. Henderson, E. Camporeale, and G. D. Reeves (2013), Modeling radiation belt electron dynamics during GEM challenge intervals with the DREAM3D diffusion model. *Journal of Geophysical Research: Space Physics*, *118*(10), 6197–6211, doi:10.1002/jgra.50560, 2013JA019063.

Walther, B. A. and J. L. Moore (2005), The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*(6), 815–829, doi:10.1111/j.2005.0906-7590.04112.x.

Wilks, D. S. (2006), *Statistical methods in the atmospheric sciences, 2nd Edition*. Academic Press.

Yu, Y., J. Koller, V. K. Jordanova, et al. (2014), Application and testing of the L* neural network with the self-consistent magnetic field model of RAM-SCB. *Journal of Geophysical Research: Space Physics*, *119*(3), 1683–1692, doi:10.1002/2013JA019350.

Zhelavskaya, I. S., M. Spasojevic, Y. Y. Shprits, and W. S. Kurth (2016), Automated determination of electron density from electric field measurements on the Van Allen Probes spacecraft. *Journal of Geophysical Research: Space Physics*, 4611–4625, doi:10.1002/2015JA022132.

Zheng, Y. and D. Rosenfeld (2015), Linear relation between convective cloud base height and updrafts and application to satellite retrievals. *Geophysical Research Letters*, *42*(15), 6485–6491, doi:10.1002/2015GL064809, 2015GL064809.